The Role of Economic Space in Decision Making: A Comment

Emmanuel Flachaire
EUREQua, Université Paris 1 Panthéon-Sorbonne

February 2004

1 Introduction

Over the last few years, Margaret Slade has contributed to some major improvments in the field of industrial economics. The important question of location and spatial interaction in economic decision is one of her central interests. Her paper, prepared for a presentation at the "*Conférence de l'ADRES*" in Paris, presents the ways and the methods she developed with her coauthors to incorporate the influence of space location in regression model. The new attention to specifying, estimating and testing for the presence of spatial interaction they have taken, concerns the use of semiparametric methods to allow less restrictions on the form of the spatial dependence. The paper is clearly written, without technical developments and the discussion of potential applications is very convincing on the significant role that the location can take in economic decisions.

In the standard linear model, there are two ways to incorporate spatial dependence: in the covariance matrix of the error term and/or in the parametric portion of the model as additional regressors. In this comment, I would like to explore different utilizations of semiparametric methods to treat the problem of location effect in regression model and compare them with the methods used by Margaret Slade and her coauthors. These different utilizations suggest that location effects could be incorporated in regression model at a low cost, with an easy estimation of the model and a simple interpretation of the estimators.

In section 2, I consider the specification of the dependence in the error term and propose to use semiparametric methods in order to obtain efficient estimators. In section 3, I consider the specification of the dependence as additional regressors and I show that the spatial dependent model can be closely linked to the semiparametric partial linear model.

2 Specification of spatially dependent model

If we consider that location can have significant consequences on economic decisions, such effects should be incorporated in the regression model. This can be done by incorporating some measures of neighborhood quality as regressors. If some indexes can correctly measure the location effects and are included as additional regressors in the model, or if the spatial dependence is correctly specified as an additional regressor in the form of a spatially lagged dependent variable, the error term is a white noise. However, neighborhood boundaries can be difficult to define and neighborhood quality can be difficult to measure. Thus, if the location effect cannot be completely specified in the parametric portion of the model, as additional regressors, the error term is dependent. Let us consider the linear model

$$y = X\beta + u, \qquad \text{with} \qquad E(u|X) = 0 \quad E(uu^\top|X) = \Omega \tag{1}$$

where the covariance matrix of the error term Ω is a non-diagonal matrix. If we have some information on the form of the dependence and if we can obtain a consistent estimator of Ω, we can use the Generalised Least Squares, or GLS, estimation method to obtain the best linear unbiased estimator for β. However, if we do not have enough information on the form of the dependence to estimate Ω consistently, the efficient GLS method cannot be used. It is then possible to use the OLS estimator of β, which is still consistent, with a covariance matrix estimator robust to heteroskedastic and dependent error term of unknown form. This last estimator is an extension of the robust estimator proposed in Newey and West (1987) to the case of spatial dependence, called *spatial Newey-West* estimator hereafter.

2.1 Spatial dependence in regressors and error term

Margaret Slade makes use of a model with a specification of the spatial dependence both in the regressors and in the error term. For instance, in section 3, she uses models that can be rewritten as

$$y_i = X_i\beta + \sum_j \lambda_{ij}\, z_j + u_i, \qquad u_i = \sum_j \rho_{ij}\, u_j + \varepsilon_i \tag{2}$$

where X_i is a row vector of k regressors, z_j is the dependent variable y_j or another variable, by convention a location is never a neighbor of itself, $\lambda_{ii} = \rho_{ii} = 0$, and ε_i is a white noise. In the estimation method (section 3) and in the application on *"measuring technological spillovers"* (section 5), the use of a semiparametric estimation of λ_{ij} *and* a spatial Newey-West estimator is recommended. The specification of the spatial dependence at the same time in the regressors, through parameters λ_{ij}, and in the error term, through parameters ρ_{ij}, suggests that the specification as additional regressors does not catch completely the spatial dependence and some of this dependence is still in the error term.

Confronted with the problem of spatial dependence, in practice, I would be tempted to adopt one of the following approaches:

1. if the spatial dependence can be correctly specified in the parametric portion of the model (through additional regressors), the error term is assumed to be independent
2. if the spatial dependence cannot be correctly specified in the regressors, the error term can be dependent and a spatial Newey-West estimator is used without any specification of the dependance in the regressors.

The use of a double specification of the spatial dependence, in the parametric portion of the model *and* in the error term, could be useful if the spatially-dependent term included as additional regressor is of primary interest. Otherwise, if the main interest concerns the fitted values of the model and thus, a valid and reliable estimator of β, a single specification in the error term can be used. It would be useful to give more evidence on the usefulness of the specification of the spatial dependence as additional regressors, when a spatial Newey-West estimator is used at the same time.

It is important to note that the Newey-West estimator should be used cautiously, because tests based on this estimator can be unreliable in finite sample. Andrews (1991) shows that this estimator can perform quite poorly in certain contexts. It follows that the use of the Newey-West estimator in practice requires very large sample size to be reliable.

2.2 Efficient estimators

The OLS parameter estimator with a spatial Newey-West covariance matrix estimator permits asymptotically correct inference on β in the presence of heteroskedasticity and spatial dependence of unknown form. This estimator is robust, but not efficient. The most efficient estimator would be obtained with the GLS estimation method, which is not feasible if we cannot obtain a consistent estimator of the covariance matrix of the error term Ω. The recent development of semiparametric methods could help us to obtain a consistent estimate of Ω and thus, to use the GLS estimation rather than the robust estimation (Newey-West).

Let us consider the model (1) with spatially dependent error term

$$u = \mathcal{R}\, u + \epsilon, \qquad \epsilon \sim \text{IID}\,(0, \sigma^2 I) \tag{3}$$

where $\mathcal{R}$ is a $n \times n$ matrix with typical component $\{\rho_{ij}\}$. Thus, we have

$$(I - \mathcal{R})\, u = \epsilon, \qquad \text{and} \qquad u = (I - \mathcal{R})^{-1}\epsilon. \tag{4}$$

where I is a $n \times n$ identity matrix. The covariance matrix of the error term u is

$$\Omega = \text{Var}(u) = E(uu^\top) = \sigma^2[(I - \mathcal{R})^\top (I - \mathcal{R})]^{-1} \tag{5}$$

If consistent estimators of σ^2 and $\mathcal{R}$ can be obtained with semiparametric methods, the efficient GLS estimation method can be used to estimate the model (1). It would be interesting to investigate the use of semiparametric methods in this way, to obtain an estimator

of β valid in the presence of spatial dependence of unknown form and efficient. This could substantially improve numerical results.

3 Spatially dependent model vs. partial linear model

In this section, I consider that the spatial dependence is correctly specified as an additional regressor in the form of a spatially lagged variable, that is, a spatial dependent model with *i.i.d.* error term. I show that this model is closely linked to a standard semiparametric model: the partial linear model. At first, I consider a simple spatial dependent model, without regressors, in the geographic context. Therefore, I study a model with regressors. Finally, I investigate the spatial model with regressors, in a product-characteristic context.

3.1 Nonparametric model

Let us consider the spatial dependent model in a geographic context,

$$y_i = \sum_j \lambda_{ij}\, y_j \; + \epsilon_i, \qquad \epsilon_i \sim \text{IID}\,(0, \sigma^2). \tag{6}$$

Margaret Slade and her coauthors develop a nonparametric estimation of the parameters λ_{ij}. They assume that the weights are defined by a common function of the distance between the two spatial locations of i and j:

$$\sum_j \lambda_{ij}\, y_j = \sum_j g(d_{ij})\, y_j \tag{7}$$

where the distance function d is a metric chosen by the practitioner, as for instance the Euclidian distance. To make the estimation possible, the weights λ_{ij} must satisfy some condition: "the influence of other locations must decay as the distance between locations increases" (section 3).

Let us compare the spatial dependent model to the following nonparametric model,

$$y_i = f(z_{1i}, z_{2i}) + \varepsilon_i, \qquad \varepsilon_i \sim \text{IID}\,(0, \sigma^2) \tag{8}$$

where z_{1i} and z_{2i} define the location of i as geographic coordinates (latitude, longitude). A nonparametric estimator of the regression function f at the point (z_{1i}, z_{2i}) can be written as a weighted sum of the dependent variable:

$$\hat{f}(z_{1i}, z_{2i}) = \sum_j w_j(z_{1i}, z_{2i})\, y_j, \tag{9}$$

where the weighting function $w_j(z_{1i}, z_{2i})$ assigns higher weights to observations close to (z_{1i}, z_{2i}), for more details see for instance Pagan and Ullah (1999, chapter 3). Many different

weighting function are candidates. Kernel estimation defines the weights with a probability function, commonly known as kernels, and a bandwidth parameter. The kernel function expresses the shape of the weights and the bandwidth parameter controls the magnitude. As a result, a large value of the bandwidth assigns greater weight to observations far from (z_{1i}, z_{2i}).

It is clear that the spatial dependent model (6) and the nonparametric model (8) are closely linked: both fitted values are written as a weighted sum of the dependent variable, with decreasing weights as the distance location increases.

3.2 Partial linear model

It is not difficult to extend the same argument to a spatial model with regressors,

$$y_i = X_i\beta + \sum_j \lambda_{ij}\, y_j + \epsilon_i, \qquad \epsilon_i \sim \text{IID}\,(0, \sigma^2) \tag{10}$$

It leads us to consider the semiparametric partial linear model

$$y_i = X_i\beta + f(z_{1i}, z_{2i}) + \varepsilon_i, \qquad \varepsilon_i \sim \text{IID}\,(0, \sigma^2) \tag{11}$$

Robinson (1988) influential paper shows that β can be estimated consistently, at a rate of convergence similar to a parametric rate. This model can be rewrite

$$y_i - E(y_i|z_{1i}, z_{2i}) = [X_i - E(X_i|z_{1i}, z_{2i})]\,\beta + \varepsilon \tag{12}$$

Robinson proposes to estimate $h_{1i} = E(y_i|z_{1i}, z_{2i})$ and $h_{2i} = E(X_i|z_{1i}, z_{2i})$ with nonparametric kernel estimators, and shows that the OLS estimator of the model

$$(y_i - \hat{h}_{1i}) = [X_i - \hat{h}_{2i}]\,\beta + \varepsilon \tag{13}$$

is a $\sqrt{n}$-consistent estimator of β, often called the "double residual" estimator. A consistent estimator of f is given by a nonparametric estimation of $y_i - X_i\hat{\beta}$ on (z_{1i}, z_{2i}),

$$\hat{f}(z_{1i}, z_{2i}) = \sum_j w_j(z_{1i}, z_{2i})\,[y_j - X_j\hat{\beta}], \tag{14}$$

Estimation of β and f requires 4 steps:

1. $\hat{h}_{1i}$ is the residual from the nonparametric estimation of y_i on (z_{1i}, z_{2i})
2. $\hat{h}_{2i}$ is the residual from the nonparametric estimation of X_i on (z_{1i}, z_{2i})
3. $\hat{\beta}$ is the OLS parameter estimator of $(y_i - \hat{h}_{1i})$ on $(X_i - \hat{h}_{2i})$
4. $\hat{f}$ is the fitted values from the nonparametric estimation of $(y_i - X_i\hat{\beta})$ on (z_{1i}, z_{2i})

Interpretation of the estimators of β and f is straightforward. In (14), we can decompose the equation in two components by developing the right term. If X_i and z_{1i}, z_{2i} are not independent and if we can write $X_i = h(z_{1i}, z_{2i}) + \eta_i$, the function h measures the influence of the location on the regressors X_i and η_i is the part of the regressors not explained by the location. A nonparametric estimator of h is given by $\hat{h}(z_{1i}, z_{2i}) = \sum_j w_j(z_{1i}, z_{2i}) X_j$. This last term is the second component of the right term in equation (14), up to scale factors. Furthermore, the first component, $\sum_j w_j(z_{1i}, z_{2i})\, y_j$, is the influence of the location on the dependent variable. This makes clear that the estimator of f measures the direct influence of the location on y_i and the indirect influence of the location on X_i. In addition, it can be shown that $\hat{\beta}$ measures the direct influence of X_i on y_i: if we replace X_i by $h(z_{1i}, z_{2i}) + \eta_i$ in (11) and if we calculate equation (12) again, the two functions f and h are removed. In other words, the influence of the location on y_i and X_i is removed when we compute an estimator of β. For more details, among others, see Yatchew (2003).

Finally, in the partial linear model, the estimator of β measures the *direct* influence of the regressors X_i on the dependent variable y_i and the estimator of f measures the influence of the location (z_{1i}, z_{2i}) on the model. The influence of the location includes at the same time a *direct* influence on the dependent variable y_i and an *indirect* influence on the regressors X_i. It follows that $\hat{\beta}$ is an estimator robust to the influence, of any form, of the location on the model.

In addition, we can see that the direct influence of the location on the dependent variable, that is, the first component in equation (14), is similar to the spatial dependent term in model (10). This makes clear that the spatial dependent model and the partial linear model are closely connected. Note that the partial linear model includes a measure of the influence of the location on the regressors, not the spatial dependent model. Therefore, it would be interesting to study further the link between these two models and to compare them based on some empirical results.

3.3 Product-Characteristic context

The previous developments are concerned with location in a geographic context. They can be applied to the spatial dependence in a product-characteristic context. Let us consider the model used by Margaret Slade in section 3.2, but with *i.i.d.* error terms, that is,

$$y_i = X_i\,\beta + \sum_j \lambda_{ij}\, p_j + \epsilon_i, \qquad \epsilon_i \sim \text{IID}\,(0, \sigma^2) \tag{15}$$

where ϵ_i is a white noise and λ_{ij} is a function of measures of distance in product-characteristic space defined by a row-vector of k variables Z_i. This spatially dependent model in prices assigns higher weights to observations that are close to i in the product-characteristic space.

For comparison purpose, let us consider the model

$$y_i = X_i\,\beta + \gamma\,p_i + \epsilon_i, \quad \text{where} \quad p_i = m(Z_i) + q_i, \tag{16}$$

where q_i is the portion of p_i not explained by Z_i. A nonparametric estimator of m is given by $\hat{m}(Z_i) = \sum_j w_j(Z_i)\,p_j$, where the weighting function assigns higher weight to observations close to i in the product-characteristic space, that is, with characteristics Z_j similar to Z_i. It is clear that this last estimator is very similar to the spatially dependent regressor in (15). It leads us to consider the following partial linear model,

$$y_i = X_i\,\beta + \gamma\,p_i + m(Z_i) + \epsilon_i, \qquad \epsilon_i \sim \text{IID}\,(0, \sigma^2). \tag{17}$$

This model includes at the same time the two equations defined in (16) and we have

$$\hat{m}(Z_i) = \sum_j w_j(Z_i)\,[y_j - X_j\,\hat{\beta} - \hat{\gamma}\,p_j] \tag{18}$$

With the same argument as in the geographic context, we can see that the estimator of γ measures the direct influence of the price on the dependent variable and $m(Z_i)$ measures the influence of the product-characteristic location Z_i on the price and on the other variables of the model.

There is some limitations to the use of the model (17) in practice, because a nonparametric estimation of the function m would be unreliable with more than three variables in Z_i, unless a huge sample is available. This problem is known as the *curse of dimensionality*. However, an usual way to reduce the number of dimensions in the nonparametric portion of the model is to use only discrete and continuous variables in the unknown function m. Indeed, dummy variables would cause only scale effects and would not affect the curvature of the function if they were included in m. Thus, the presence of dummy variables in the product-characteristic space will not be included in the nonparametric part of the model but as regressors in the parametric part of the model. This contributes to reduce the curse of dimensionality.

Once more, we can see that the spatial dependent model and the partial linear model are closely linked and it would be interesting to compare numerical empirical results based on these two models.

4 Conclusion

In this comment, I have explored the use of semiparametric methods to incorporate the effects of spatial location in regression model, in a different way that the methods developped by Margaret Slade and her coauthors. On the one hand, if our interest is mainly concerned by the estimation of a model, robust to the influence of the location, we have seen that semiparametric methods could be used to obtain efficient estimators. On the other hand, if our main interest is to measure the influence of the location, we have seen that the

spatial dependent model is closely connected to the partial linear model. I have presented some similarities between the spatial dependent model and the partial linear model, further developments should study their differences.

References

Andrews, D. W. K. (1991). "Heteroskedasticity and autocorrelation consistent covariance matrix estimation". *Econometrica 59*, 817–858.

Newey, W. K. and K. D. West (1987). "A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix". *Econometrica 55*, 703–708.

Pagan, A. and A. Ullah (1999). *Nonparametric Econometrics*. Cambridge University Press, Cambridge.

Robinson, P. M. (1988). "Root-n-consistent semiparametric regression". *Econometrica 56*, 931–954.

Yatchew, A. (2003). *Semiparametric Regression for the Applied Econometrician*. Cambridge University Press, Cambridge.